Procédé et système de conversion rapides d'un signal vocal.

La présente invention concerne un procédé de conversion d'un signal vocal prononcé par un locuteur source en un signal vocal converti dont les caractéristiques acoustiques ressemblent à celles d'un locuteur cible ainsi qu'un système mettant en œuvre un tel procédé.

Dans le cadre d'applications de conversion de voix, telles que les services vocaux, les applications de dialogue oral homme-machine ou encore la synthèse vocale de textes, le rendu auditif est primordial et, pour obtenir une qualité acceptable, il convient de bien maîtriser les paramètres liés à la prosodie des signaux vocaux.

De manière classique, les principaux paramètres acoustiques ou prosodiques modifiés lors de procédés de conversion de voix sont les paramètres relatifs à l'enveloppe spectrale et/ou pour les sons voisés faisant intervenir la vibration des cordes vocales, les paramètres relatifs à une structure périodique, soit la période fondamentale dont l'inverse est appelé fréquence fondamentale ou « pitch ».

Les procédés de conversion de voix classiques comprennent en général la détermination d'au moins une fonction de transformation de caractéristiques acoustiques du locuteur source en caractéristiques acoustiques proches de celles du locuteur cible, et la transformation d'un signal vocal à convertir par l'application de cette ou ces fonctions.

Cette transformation est une opération longue et coûteuse en temps de calcul.

En effet, de telles fonctions de transformation sont classiquement considérées comme des combinaisons linéaires d'un nombre fini important d'éléments de transformation appliqués à des éléments représentatifs du signal vocal à convertir.

Le but de l'invention est de résoudre ces problèmes en définissant un procédé et un système de conversion d'un signal vocal rapide et de bonne qualité.

A cet effet, la présente invention a pour objet un procédé de conversion d'un signal vocal prononcé par un locuteur source en un signal vocal converti dont les caractéristiques acoustiques ressemblent à celles d'un locuteur cible, comprenant :

- la détermination d'au moins une fonction de transformation de caractéristiques acoustiques du locuteur source en caractéristiques acoustiques proches de celles du locuteur cible, à partir d'échantillons vocaux des locuteurs source et cible ; et

- la transformation de caractéristiques acoustiques du signal vocal à convertir du locuteur source, par l'application de ladite au moins une fonction de transformation,

caractérisé en ce que ladite transformation comprend une étape d'application uniquement d'une partie déterminée d'au moins une fonction de transformation sur ledit signal à convertir.

Le procédé de l'invention permet ainsi de diminuer le temps de calcul nécessaire à la mise en œuvre, grâce à l'application uniquement d'une partie déterminée d'au moins une fonction de transformation.

Suivant d'autres caractéristiques de l'invention :

- au moins la détermination d'une fonction de transformation comprend une étape de détermination d'un modèle représentant de manière pondérée des caractéristiques acoustiques communes des échantillons vocaux du locuteur cible et du locuteur source sur un ensemble fini de composantes de modèle, et ladite transformation comprend :

- une étape d'analyse du signal vocal à convertir, regroupé en trames pour obtenir, pour chaque trame d'échantillons des informations relatives aux caractéristiques acoustiques ;

- une étape de détermination d'un indice de correspondance entre les trames à convertir et chaque composante dudit modèle ; et

- une étape de sélection d'une partie déterminée desdites composantes dudit modèle en fonction desdits indices de correspondance,

ladite étape d'application uniquement d'une partie déterminée d'au moins une fonction de transformation comprenant l'application auxdites trames à convertir de la seule partie de ladite au moins une fonction de transformation correspondant auxdites composantes du modèle sélectionnées ;

- il comporte en outre une étape de normalisation de chacun desdits indices de correspondance des composantes sélectionnées par rapport à la somme de tous les indices de correspondance des composantes sélectionnées ;

- il comporte en outre une étape de mémorisation desdits indices de correspondance et de ladite partie déterminée desdites composantes de modèle, réalisée avant ladite étape de transformation, laquelle est retardée dans le temps ;

- ladite détermination de ladite au moins une fonction de transformation comprend :

- une étape d'analyse des échantillons vocaux des locuteurs source et cible, regroupés en trame pour obtenir des caractéristiques acoustiques pour chaque trame d'échantillons d'un locuteur ;

- une étape d'alignement temporel des caractéristiques acoustiques du locuteur source avec les caractéristiques acoustiques du locuteur cible, cette étape étant réalisée avant ladite étape de détermination d'un modèle ;

- ladite étape de détermination d'un modèle correspond à la détermination d'un modèle de mélange de densités de probabilités gaussiennes ;

- ladite étape de détermination d'un modèle comprend :

- une sous-étape de détermination d'un modèle correspondant à un mélange de densités de probabilités gaussiennes, et

- une sous-étape d'estimation des paramètres du mélange de densités de probabilités gaussiennes à partir de l'estimation du maximum de vraisemblance entre les caractéristiques acoustiques des échantillons des locuteurs source et cible et le modèle ;

- ladite détermination d'au moins une fonction de transformation est réalisée à partir d'un estimateur de la réalisation des caractéristiques acoustiques du locuteur cible sachant les caractéristiques acoustiques du locuteur source ;

- ledit estimateur est formé de l'espérance conditionnelle de la réalisation des caractéristiques acoustiques du locuteur cible sachant la réalisation des caractéristiques acoustiques du locuteur source ;

- il comporte en outre une étape de synthèse permettant de former un signal vocal converti à partir desdites informations acoustiques transformées.

L'invention à également pour objet un système de conversion d'un signal vocal prononcé par un locuteur source en un signal vocal converti dont les caractéristiques acoustiques ressemblent à celles d'un locuteur cible, comprenant :

- des moyens de détermination d'au moins une fonction de transformation des caractéristiques acoustiques du locuteur source en caractéristiques acoustiques proches de celles du locuteur cible, à partir d'échantillons vocaux des locuteurs source et cible ; et

- des moyens de transformation des caractéristiques acoustiques du signal vocal à convertir du locuteur source par l'application de ladite au moins une fonction de transformation,

caractérisé en ce que lesdits moyens de transformation sont adaptés pour l'application uniquement d'une partie déterminée d'au moins une fonction de transformation sur ledit signal à convertir.

Selon d'autres caractéristiques du système :

- lesdits moyens de détermination sont adaptés pour la détermination d'au moins une fonction de transformation à l'aide d'un modèle représentant de manière pondérée des caractéristiques acoustiques communes des échantillons vocaux des locuteurs source et cible sur un ensemble fini de composantes, et en ce qu'il comporte :

- des moyens d'analyse dudit signal à convertir, regroupé en trames, pour obtenir, pour chaque trame d'échantillons, des informations relatives aux caractéristiques acoustiques ;

- des moyens de détermination d'un indice de correspondance entre les trames à convertir et chaque composante dudit modèle ; et

- des moyens de sélection d'une partie déterminée desdites composantes dudit modèle en fonction desdits indices de correspondance,

lesdits moyens d'application étant adaptés pour appliquer uniquement une partie déterminée de ladite au moins une fonction de transformation correspondant auxdites composantes du modèle sélectionnées.

L'invention sera mieux comprise à la lecture de la description qui va suivre, donnée uniquement à titre d'exemple et faite en se référant aux dessins annexés, sur lesquels :

- les Figs. 1A et 1B représentent un organigramme général du procédé de l'invention ; et

- la Fig. 2 représente un schéma synoptique d'un système mettant en œuvre le procédé de l'invention.

La conversion de voix consiste à modifier le signal vocal d'un locuteur de référence appelé locuteur source, de telle sorte que le signal produit semble avoir été prononcé par un autre locuteur, nommé locuteur cible.

Un tel procédé comporte tout d'abord la détermination de fonctions de transformation de caractéristiques acoustiques ou prosodiques, des signaux vocaux du locuteur source en caractéristiques acoustiques proches de celles des signaux vocaux du locuteur cible, à partir d'échantillons vocaux prononcés par le locuteur source et le locuteur cible.

Plus particulièrement, la détermination 1 de fonctions de transformation est réalisée sur des bases de données d'échantillons vocaux correspondant à la réalisation acoustique de mêmes séquences phonétiques, prononcées respectivement par les locuteurs source et cible.

Cette détermination est désignée sur la figure 1A par la référence numérique générale 1 et est également couramment appelée « apprentissage ».

Le procédé comporte ensuite une transformation des caractéristiques acoustiques d'un signal vocal à convertir prononcé par le locuteur source à l'aide de la ou des fonctions déterminées précédemment. Cette transformation est désignée par la référence numérique générale 2 sur la figure 1B.

En fonction des modes de réalisation, différentes caractéristiques acoustiques sont transformées telles que des caractéristiques d'enveloppe spectrale et/ou de fréquence fondamentale.

Le procédé débute par des étapes 4X et 4Y d'analyse d'échantillons vocaux prononcés respectivement par les locuteurs source et cible. Ces étapes permettent de regrouper les échantillons par trame, afin d'obtenir pour chaque trame d'échantillons, des informations relatives à l'enveloppe spectrale et/ou des informations relatives à la fréquence fondamentale.

Dans le mode de réalisation décrit, les étapes 4X et 4Y d'analyse sont fondées sur l'utilisation d'un modèle de signal sonore sous la forme d'une somme d'un signal harmonique avec un signal de bruit selon un modèle communément appelé "HNM" (en anglais : Harmonic plus Noise Model).

Le modèle HNM comprend la modélisation de chaque trame de signal vocal en une partie harmonique représentant la composante périodique du signal, constituée d'une somme de L sinusoïdes harmoniques d'amplitude A_l et de

phase ϕ_l, et d'une partie bruitée représentant le bruit de friction et la variation de l'excitation glottale.

On peut ainsi écrire :

$s(n)=h(n)+b(n)$

avec $h(n)= \sum_{l=1}^{L} A_l(n)\cos(\phi_l(n))$

Le terme h(n) représente donc l'approximation harmonique du signal s(n).

En outre, le mode de réalisation décrit est fondé sur une représentation de l'enveloppe spectrale par le cepstre discret.

Les étapes 4X et 4Y comportent des sous-étapes 8X et 8Y d'estimation, pour chaque trame, de la fréquence fondamentale, par exemple au moyen d'une méthode d'auto corrélation.

Les sous-étapes 8X et 8Y sont chacune suivies d'une sous-étape 10X et 10Y d'analyse synchronisée de chaque trame sur sa fréquence fondamentale, qui permet d'estimer les paramètres de la partie harmonique ainsi que les paramètres du bruit du signal et notamment la fréquence maximale de voisement. En variante, cette fréquence peut être fixée arbitrairement ou être estimée par d'autres moyens connus.

Dans le mode de réalisation décrit, cette analyse synchronisée correspond à la détermination des paramètres des harmoniques par minimisation d'un critère de moindres carrés pondérés entre le signal complet et sa décomposition harmonique correspondant dans le mode de réalisation décrit, au signal de bruit estimé. Le critère noté E est égal à :

$$E = \sum_{n=-T_i}^{T_i} w^2(n)(s(n)-h(n))^2$$

Dans cette équation, w (n) est la fenêtre d'analyse et T_i est la période fondamentale de la trame courante.

Ainsi, la fenêtre d'analyse est centrée autour de la marque de la période fondamentale et a pour durée deux fois cette période.

En variante, ces analyses sont faites de manière asynchrone avec un pas fixe d'analyse et une fenêtre de taille fixe.

Les étapes 4X et 4Y d'analyse comportent enfin des sous-étapes 12X et 12Y d'estimation des paramètres de l'enveloppe spectrale des signaux en utili-

sant par exemple une méthode de cepstre discret régularisé et une transformation en échelle de Bark pour reproduire le plus fidèlement possible les propriétés de l'oreille humaine.

Ainsi, les étapes 4X et 4Y d'analyse délivrent respectivement pour les échantillons vocaux prononcés par les locuteurs source et cible, pour chaque trame de rang n d'échantillons des signaux de parole, un scalaire noté F_n représentant la fréquence fondamentale et un vecteur noté c_n comprenant des informations d'enveloppe spectrale sous la forme d'une séquence de coefficients cepstraux.

Le mode de calcul des coefficients cepstraux correspond à un mode opératoire connu de l'état de la technique et, pour cette raison, ne sera pas décrit plus en détail.

Le procédé de l'invention permet donc de définir pour chaque trame n du locuteur source, un vecteur noté x_n de coefficients cepstraux $c_x(n)$ et la fréquence fondamentale.

De manière similaire, le procédé permet de définir pour chaque trame n de locuteur cible, un vecteur y_n de coefficients cepstraux $c_y(n)$, ainsi que la fréquence fondamentale.

Les étapes 4X et 4Y sont suivies d'une étape 18 d'alignement entre le vecteur source x_n et le vecteur cible y_n, de manière à former un appariement entre ces vecteurs obtenu par un algorithme classique d'alignement temporel dynamique dit « DTW » (en anglais : Dynamic Time Warping).

L'étape 18 d'alignement est suivie d'une étape 20 de détermination d'un modèle représentant de manière pondérée les caractéristiques acoustiques communes du locuteur source et du locuteur cible sur un ensemble fini de composantes de modèle.

Dans le mode de réalisation décrit, il s'agit d'un modèle probabiliste des caractéristiques acoustiques du locuteur cible et du locuteur source, selon un modèle noté « GMM » de mélanges de composantes formées de densités gaussiennes. Les paramètres des composantes sont estimés à partir des vecteurs source et cible contenant, pour chaque locuteur, le cepstre discret.

De manière classique, la densité de probabilité d'une variable aléatoire notée de manière générale p(z), suivant un modèle de mélange de densités de

probabilités gaussiennes GMM s'écrit mathématiquement de la manière suivante :

$$p(z)=\sum_{i=1}^{Q}\alpha_i x N(z,\mu_i,\Sigma_i)$$

avec $$\sum_{i=1}^{Q}\alpha_i = 1,\ o\leq\alpha_i\leq 1$$

Dans cette formule, Q désigne le nombre de composantes du modèle, $N(z ; \mu_i, \Sigma_i)$ est la densité de probabilité de la loi normale de moyenne μ_i et de matrice de covariance Σ_i et les coefficients α_i sont les coefficients du mélange.

Ainsi, le coefficient α_i correspond à la probabilité a priori que la variable aléatoire z soit générée par la $i^{ème}$ composante gaussienne du mélange.

De manière plus particulière, l'étape 20 de détermination du modèle comporte une sous-étape 22 de modélisation de la densité jointe p(z) des vecteurs source noté x et cible noté y, de sorte que :

$$Z_n = \left[x_n^T , y_n^T\right]^T$$

L'étape 20 comporte ensuite une sous-étape 24 d'estimation de paramètres GMM (α, μ, Σ) de la densité p(z). Cette estimation peut être réalisée, par exemple, à l'aide d'un algorithme classique de type dit "EM" (Expectation – Maximisation), correspondant à une méthode itérative conduisant à l'obtention d'un estimateur de maximum de vraisemblance entre les données des échantillons de parole et le modèle de mélange de gaussiennes.

La détermination des paramètres initiaux du modèle GMM est obtenue à l'aide d'une technique classique de quantification vectorielle.

L'étape 20 de détermination de modèle délivre ainsi les paramètres d'un mélange de densités gaussiennes représentatifs des caractéristiques acoustiques communes des échantillons vocaux du locuteur source et du locuteur cible.

Le modèle ainsi défini forme donc une représentation pondérée de caractéristiques acoustiques d'enveloppe spectrale communes des échantillons vocaux du locuteur cible et du locuteur source sur l'ensemble fini de composantes du modèle.

Le procédé comporte ensuite une étape 30 de détermination, à partir du modèle et des échantillons vocaux, d'une fonction de transformation de l'enveloppe spectrale du signal du locuteur source vers le locuteur cible.

Cette fonction de transformation est déterminée à partir d'un estimateur de la réalisation des caractéristiques acoustiques du locuteur cible étant donné les caractéristiques acoustiques du locuteur source, formé dans le mode de réalisation décrit, par l'espérance conditionnelle.

Pour cela, l'étape 30 comporte une sous-étape 32 de détermination de l'espérance conditionnelle des caractéristiques acoustiques du locuteur cible sachant les informations caractéristiques acoustiques du locuteur source. L'espérance conditionnelle est notée F(x) et est déterminée à partir des formules suivantes :

$$F(x)=E[y \mid x]=\sum_{i=1}^{Q} h_i(x)[\mu_i^y + \Sigma_i^{yx}(\Sigma_i^{xx})^{-1}(x-\mu_i^x)]$$

avec $$h_i(x)=\frac{\alpha_i N(x,\mu_i^x,\Sigma_i^{xx})}{\sum_{j=1}^{Q}\alpha_i N(x,\mu_j^x,\Sigma_j^{xx})}$$

avec $$\Sigma_i=\begin{bmatrix}\Sigma_i^{xx} & \Sigma_i^{xy}\\ \Sigma_i^{yx} & \Sigma_i^{yy}\end{bmatrix} \text{ et } \mu_i=\begin{bmatrix}\mu_i^x\\ \mu_i^y\end{bmatrix}$$

Dans ces équations, $h_i(x)$ correspond à la probabilité a posteriori que le vecteur source x soit généré par la $i^{ème}$ composante du modèle de mélange de densités gaussiennes du modèle, et le terme entre crochets correspond à un élément de transformation déterminé à partir du modèle. On rappelle que y désigne le vecteur cible.

La détermination de l'espérance conditionnelle permet ainsi d'obtenir la fonction de transformation des caractéristiques d'enveloppe spectrale entre le locuteur source et le locuteur cible sous la forme d'une combinaison linéaire pondérée d'éléments de transformation.

L'étape 30 comporte également une sous-étape 34 de détermination d'une fonction de transformation de la fréquence fondamentale par une mise à l'échelle de la fréquence fondamentale du locuteur source, sur la fréquence fondamentale du locuteur cible. Cette étape 34 est réalisée de manière classique à

un instant quelconque du procédé à l'issue des sous-étapes 8X et 8Y d'estimation de la fréquence fondamentale.

En référence à la figure 1B, le procédé de conversion comporte ensuite la transformation 2 d'un signal vocal à convertir prononcé par le locuteur source, lequel signal à convertir peut être différent des signaux vocaux utilisés précédemment.

Cette transformation 2 débute par une étape d'analyse 36 réalisée, dans le mode de réalisation décrit, à l'aide d'une décomposition selon le modèle HNM similaire à celles réalisées dans les étapes 4X et 4Y décrites précédemment. Cette étape 36 permet de délivrer des informations d'enveloppe spectrale sous la forme de coefficients cepstraux, des informations de fréquence fondamentale ainsi que des informations de phase et de fréquence maximale de voisement.

Cette étape 36 d'analyse est suivie d'une étape 38 de détermination d'un indice de correspondance entre le vecteur à convertir et chaque composante du modèle.

Dans le mode de réalisation décrit, chacun de ces indices correspond à la probabilité a posteriori de la réalisation du vecteur à convertir par chacune des différentes composantes du modèle, soit au terme $h_i(x)$.

Le procédé comporte ensuite une étape 40 de sélection d'un nombre restreint de composantes du modèle en fonction des indices de correspondance déterminés à l'étape précédente, lequel ensemble restreint est noté S(x).

Cette étape 40 de sélection est mise en œuvre par une procédure itérative permettant de retenir un ensemble minimal de composantes, ces composantes étant sélectionnées tant que la somme cumulée de leurs indices de correspondance est inférieure à un seuil prédéterminé.

En variante, cette étape de sélection comprend la sélection d'un nombre fixe de composantes dont les indices de correspondance sont les plus élevés.

Dans le mode de réalisation décrit, l'étape 40 de sélection est suivie d'une étape 42 de normalisation des indices de correspondance des composantes sélectionnées du modèle. Cette normalisation est réalisée par le rapport de chaque indice sélectionné sur la somme de tous les indices sélectionnés.

Avantageusement, le procédé comporte ensuite une étape 43 de stockage des composantes de modèle sélectionnées ainsi que des indices de correspondance normalisés associés.

Une telle étape 43 de mémorisation est particulièrement utile dans le cas où l'analyse est réalisée en temps différé par rapport au reste de la transformation 2, qui permet de préparer efficacement une conversion ultérieure.

Le procédé comporte ensuite une étape 44 d'application partielle de la fonction de transformation de l'enveloppe spectrale par l'application des seuls éléments de transformation correspondant aux composantes de modèle sélectionnées. Ces seuls éléments de transformation sélectionnés sont appliqués aux trames du signal à convertir, afin de réduire le temps nécessaire à la mise en œuvre de cette transformation.

Cette étape 44 d'application correspond à la résolution de l'équation suivante pour les seules composantes sélectionnées de modèle formant l'ensemble restant S(x), de sorte que

$$F(x) = \sum_{i \in S(x)} w_i(x)[\mu_i^y + \Sigma_i^{yx}(\Sigma_i^{xx})^{-1}(x-\mu_i^x)]$$

avec $$w_i(x) = \frac{h_i(x)}{\sum_{i \in S(x)} h_i(x)}$$

Ainsi, pour une trame donnée, avec p la dimension des vecteurs de données, Q le nombre total de composantes et N le nombre de composantes sélectionnées, l'étape 44 d'application partielle de la fonction de transformation se limite à N (P^2 + 1) multiplications, qui se rajoutent aux Q (P^2 + 1) modifications permettant de déterminer les indices de correspondance, contre deux fois $Q(P^2+1)$. En conséquence, la réduction de complexité obtenue est au moins de l'ordre de Q/(Q+N).

De plus, dans le cas où le résultat des étapes 36 à 42 a été mémorisé, grâce à la réalisation de l'étape 43, l'étape 44 d'application de la fonction de transformation se limite à $N(P^2+1)$ opérations contre $2Q(P^2+1)$, dans l'état de la technique, de sorte que, pour cette étape 44, la réduction du temps de calcul est de l'ordre de 2Q/N.

La qualité de la transformation est cependant préservée par l'application des composantes présentant un indice de correspondance élevé avec le signal à convertir.

Le procédé comporte ensuite une étape 46 de transformation des caractéristiques de fréquence fondamentale du signal vocal à convertir, à l'aide de la fonction de transformation par mise à l'échelle déterminée à l'étape 34 et réalisée selon des techniques classiques.

De manière également classique, le procédé de conversion comporte ensuite une étape 48 de synthèse du signal de sortie réalisée, dans l'exemple décrit, par une synthèse de type HNM qui délivre directement le signal vocal converti à partir des informations d'enveloppe spectrale transformées à l'étape 44 et des informations de fréquence fondamentale délivrées par l'étape 46. Cette étape 48 utilise également des informations de phase et de fréquence maximale de voisement délivrées par l'étape 36.

Le procédé de conversion de l'invention permet ainsi de réaliser une conversion de haute qualité avec une faible complexité et donc un gain de temps de calcul important.

Sur la figure 2, on a représenté un schéma synoptique d'un système de conversion de voix mettant en œuvre le procédé décrit en référence aux figures 1A et 1B.

Ce système utilise en entrée une base de données 50 d'échantillons vocaux prononcés par le locuteur source et une base de données 52 contenant au moins les mêmes échantillons vocaux prononcés par le locuteur cible.

Ces deux bases de données sont utilisées par un module 54 de détermination de fonctions de transformation de caractéristiques acoustiques et du locuteur source en caractéristiques acoustiques du locuteur cible.

Ce module 54 est adapté pour la mise en œuvre de l'étape 1 telle que décrite en référence à la figure 1 et permet donc la détermination d'au moins une fonction de transformation de caractéristiques acoustiques et notamment la fonction de transformation des caractéristiques d'enveloppe spectrale et la fonction de transformation de la fréquence fondamentale.

Notamment, le module 54 est adapté pour la détermination de la fonction de transformation de l'enveloppe spectrale à partir d'un modèle représentant de manière pondérée des caractéristiques acoustiques communes des échantil-

lons vocaux du locuteur cible et du locuteur source, sur un ensemble fini de composantes de modèles.

Le système de conversion de voix reçoit en entrée un signal vocal 60 correspondant à un signal de parole prononcé par le locuteur source et destiné à être converti.

Le signal 60 est introduit dans un module 62 d'analyse mettant en œuvre, par exemple une décomposition de type HNM permettant d'extraire des informations d'enveloppe spectrale du signal 60 sous la forme de coefficients cepstraux et des informations de fréquence fondamentale. Le module 62 délivre également des informations de phase et de fréquence maximales de voisement obtenues par l'application du modèle HNM.

Le module 62 met donc en œuvre l'étape 36 du procédé tel décrit précédemment.

Eventuellement, le module 62 est mis en œuvre au préalable et les informations sont stockées pour être utilisées ultérieurement.

Le système comporte ensuite un module 64 de détermination des indices de correspondance entre le signal vocal à convertir 60 et chaque composante du modèle. A cet effet, le module 64 reçoit les paramètres du modèle déterminé par le module 54.

Le module 64 met donc en œuvre l'étape 38 du procédé tel que décrit précédemment.

Le système comprend ensuite un modèle 65 de sélection de composantes du modèle mettant en œuvre l'étape 40 de procédé décrit précédemment et permettant la sélection de composantes présentant un indice de correspondance traduisant une forte connexité avec le signal vocal à convertir.

Avantageusement, ce module 65 réalise également la normalisation des indices de correspondance des composantes sélectionnées par rapport à leur moyenne en mettant en œuvre l'étape 42.

Le procédé comporte ensuite un module 66 d'application partielle de la fonction de transformation de l'enveloppe spectrale déterminée par le module 54, par l'application des seuls éléments de transformation sélectionnés par le module 65 en fonction des indices de correspondance.

Ainsi, ce module 66 est adapté pour la mise en œuvre de l'étape 44 d'application partielle de la fonction de transformation, de manière à délivrer en

sortie, des informations acoustiques du locuteur source transformées par les seuls éléments sélectionnés de la fonction de transformation, soit par les composantes du modèle présentant un indice de correspondance élevé, avec les trames du signal à convertir 60. Ce module permet donc une transformation rapide du signal vocal à convertir grâce à l'application partielle de la fonction de transformation.

La qualité de la transformation est préservée par la sélection des composantes du modèle présentant un indice élevé de correspondance avec le signal à convertir.

Le module 66 est également adapté pour réaliser une transformation des caractéristiques de fréquence fondamentale, réalisée de manière classique par l'application de la fonction de transformation par mise à l'échelle réalisée selon l'étape 46.

Le système comporte ensuite un module 68 de synthèse recevant en entrée, les informations d'enveloppe spectrale et de fréquence fondamentale transformées et délivrées par le module 66 ainsi que des informations de phase et de fréquence maximale de voisement délivrées par le module 62 d'analyse.

Le module 68 met ainsi en œuvre l'étape 46 du procédé décrit en référence à la figure 1 et délivre un signal 70, correspondant au signal vocal 60 du locuteur source mais dont les caractéristiques d'enveloppe spectrale et de fréquence fondamentale, ont été modifiées afin d'être similaires à celles du locuteur cible.

Le système décrit peut être mis en œuvre de diverses manières et notamment à l'aide de programmes informatiques adaptés et reliés à des moyens matériels d'acquisition sonore.

Ce système peut également être mis en œuvre sur des bases de données déterminées afin de former des bases de données de signaux convertis prêts à être utilisés.

Notamment, ce système peut être mis en œuvre dans une première phase de fonctionnement afin de délivrer, pour une base de données de signaux, des informations relatives aux composantes du modèle sélectionnées ainsi qu'à leurs indices de correspondance respectifs, ces informations étant alors mémorisées.

Les modules 66 et 68 du système, sont mis en œuvre ultérieurement à la demande, pour générer un signal vocal de synthèse en utilisant les signaux vocaux à convertir et les informations relatives aux composantes sélectionnées et à leurs indices de correspondance afin d'obtenir une réduction maximale du temps de calcul.

En fonction de la complexité des signaux et de la qualité souhaitée, le procédé de l'invention et le système correspondant peuvent également être mis en œuvre en temps réel.

En variante, le procédé de l'invention et le système correspondant sont adaptés pour la détermination de plusieurs fonctions de transformation. Par exemple, une première et seconde fonctions sont déterminées pour la transformation respectivement des paramètres d'enveloppe spectrale et des paramètres de fréquence fondamentale des trames à caractère voisé et une troisième fonction est déterminée pour la transformation des trames à caractère non voisé.

Dans un tel mode de réalisation, il est donc prévu une étape de séparation, dans le signal vocal à convertir, des trames voisées et non voisées et une ou plusieurs étapes de transformation de chacun de ces ensembles de trames.

Dans le cadre de l'invention, une seule ou plusieurs des fonctions de transformation est appliquée partiellement de manière à diminuer le temps de traitement.

Par ailleurs, dans l'exemple décrit, la conversion de voix est réalisée par transformation des caractéristiques d'enveloppe spectrale et des caractéristiques de fréquence fondamentale de manière séparée, seule la fonction de transformation de l'enveloppe spectrale étant appliquée partiellement. En variante, plusieurs fonctions de transformation de différentes caractéristiques acoustiques et/ou de transformation simultanées de plusieurs caractéristiques acoustiques sont déterminées et au moins l'une de ces fonctions de transformation est appliquée partiellement.

De manière générale, le système est adapté pour la mise en œuvre de toutes les étapes du procédé décrit en référence aux figures 1A et 1B.

Bien entendu, d'autres modes de réalisation que ceux décrits, peuvent être envisagés.

Notamment, les modèles HNM et GMM peuvent être remplacés par d'autres techniques et modèles connus de l'homme de l'art. Par exemple,

l'analyse est réalisée à l'aide de techniques dites LPC (Linear Predictive Coding), de modèles sinusoïdaux ou MBE (Multi Band Excited), les paramètres spectraux sont des paramètres dits LSF (Line Spectrum Frequencies), ou encore des paramètres liés aux formants ou à un signal glottique. En variante, le modèle GMM est remplacé par une quantification vectorielle floue (Fuzzy VQ.).

En variante, l'estimateur mis en œuvre lors de l'étape 30 peut être un critère de maximum a posteriori, dit "MAP" et correspondant à la réalisation du calcul de l'espérance uniquement pour le modèle représentant le mieux le couple de vecteurs source-cible.

Dans une autre variante, la détermination d'une fonction de transformation est réalisée à l'aide d'une technique dite des moindres carrés au lieu de l'estimation de la densité jointe décrite.

Dans cette variante, la détermination d'une fonction de transformation comprend la modélisation de la densité de probabilité des vecteurs source à l'aide d'un modèle GMM puis la détermination des paramètres du modèle à l'aide d'un algorithme EM. La modélisation prend ainsi en compte des segments de parole du locuteur source dont les correspondants prononcés par le locuteur cible ne sont pas disponibles.

La détermination comprend ensuite la minimisation d'un critère des moindres carrés entre paramètres cible et source pour obtenir la fonction de transformation. Il est à noter que l'estimateur de cette fonction s'exprime toujours de la même manière mais que les paramètres sont estimés différemment et que des données supplémentaires sont prises en compte.

REVENDICATIONS

1. Procédé de conversion d'un signal vocal (60) prononcé par un locuteur source en un signal vocal converti (70) dont les caractéristiques acoustiques ressemblent à celles d'un locuteur cible, comprenant :

- la détermination (1) d'au moins une fonction de transformation de caractéristiques acoustiques du locuteur source en caractéristiques acoustiques proches de celles du locuteur cible, à partir d'échantillons vocaux des locuteurs source et cible ; et

- la transformation (2) de caractéristiques acoustiques du signal vocal à convertir du locuteur source, par l'application de ladite au moins une fonction de transformation,

caractérisé en ce que ladite transformation (2) comprend une étape (44) d'application uniquement d'une partie déterminée d'au moins une fonction de transformation sur ledit signal à convertir.

2. Procédé selon la revendication 1, caractérisé en ce qu'au moins la détermination (1) d'une fonction de transformation comprend une étape (20) de détermination d'un modèle représentant de manière pondérée des caractéristiques acoustiques communes des échantillons vocaux du locuteur cible et du locuteur source sur un ensemble fini de composantes de modèle, et en ce que ladite transformation (2) comprend :

- une étape (36) d'analyse du signal vocal à convertir, regroupé en trames pour obtenir, pour chaque trame d'échantillons des informations relatives aux caractéristiques acoustiques ;

- une étape (38) de détermination d'un indice de correspondance entre les trames à convertir et chaque composante dudit modèle ; et

- une étape (40) de sélection d'une partie déterminée desdites composantes dudit modèle en fonction desdits indices de correspondance,

ladite étape (44) d'application uniquement d'une partie déterminée d'au moins une fonction de transformation comprenant l'application auxdites trames à convertir de la seule partie de ladite au moins une fonction de transformation correspondant auxdites composantes du modèle sélectionnées.

3. Procédé selon la revendication 2, caractérisé en ce qu'il comporte en outre une étape (42) de normalisation de chacun desdits indices de corres-

pondance des composantes sélectionnées par rapport à la somme de tous les indices de correspondance des composantes sélectionnées.

4. Procédé selon l'une quelconque des revendications 2 et 3, caractérisé en ce qu'il comporte en outre une étape (43) de mémorisation desdits indices de correspondance et de ladite partie déterminée desdites composantes de modèle, réalisée avant ladite étape (44) de transformation, laquelle est retardée dans le temps.

5. Procédé selon l'une quelconque des revendications 2 à 4, caractérisé en ce que ladite détermination (1) de ladite au moins une fonction de transformation comprend :

- une étape (4X, 4Y) d'analyse des échantillons vocaux des locuteurs source et cible, regroupés en trame pour obtenir des caractéristiques acoustiques pour chaque trame d'échantillons d'un locuteur ;

- une étape (18) d'alignement temporel des caractéristiques acoustiques du locuteur source avec les caractéristiques acoustiques du locuteur cible, cette étape (18) étant réalisée avant ladite étape (20) de détermination d'un modèle.

6. Procédé selon l'une quelconque des revendications 2 à 4, caractérisé en ce que ladite étape (20) de détermination d'un modèle correspond à la détermination d'un modèle de mélange de densités de probabilités gaussiennes.

7. Procédé selon la revendication 6, caractérisé en ce que ladite étape de détermination (20) d'un modèle comprend :

- une sous-étape (22) de détermination d'un modèle correspondant à un mélange de densités de probabilités gaussiennes, et

- une sous-étape (24) d'estimation des paramètres du mélange de densités de probabilités gaussiennes à partir de l'estimation du maximum de vraisemblance entre les caractéristiques acoustiques des échantillons des locuteurs source et cible et le modèle.

8. Procédé selon l'une quelconque des revendications 1 à 7, caractérisé en ce que ladite détermination (1) d'au moins une fonction de transformation est réalisée à partir d'un estimateur de la réalisation des caractéristiques acoustiques du locuteur cible sachant les caractéristiques acoustiques du locuteur source.

9. Procédé selon la revendication 8, caractérisé en ce que ledit estimateur est formé de l'espérance conditionnelle de la réalisation des caractéristiques acoustiques du locuteur cible sachant la réalisation des caractéristiques acoustiques du locuteur source.

10. Procédé selon l'une quelconque des revendications 1 à 9, caractérisé en ce qu'il comporte en outre une étape (48) de synthèse permettant de former un signal vocal converti à partir desdites informations acoustiques transformées.

11. Système de conversion d'un signal vocal (60) prononcé par un locuteur source en un signal vocal converti (70) dont les caractéristiques acoustiques ressemblent à celles d'un locuteur cible, comprenant :

- des moyens (56) de détermination d'au moins une fonction de transformation des caractéristiques acoustiques du locuteur source en caractéristiques acoustiques proches de celles du locuteur cible, à partir d'échantillons vocaux des locuteurs source et cible ; et

- des moyens (66) de transformation des caractéristiques acoustiques du signal vocal à convertir (60) du locuteur source par l'application de ladite au moins une fonction de transformation,

caractérisé en ce que lesdits moyens (66) de transformation sont adaptés pour l'application uniquement d'une partie déterminée d'au moins une fonction de transformation sur ledit signal à convertir (60).

12. Système selon la revendication 11, caractérisé en ce que lesdits moyens (54) de détermination sont adaptés pour la détermination d'au moins une fonction de transformation à l'aide d'un modèle représentant de manière pondérée des caractéristiques acoustiques communes des échantillons vocaux des locuteurs source et cible sur un ensemble fini de composantes, et en ce qu'il comporte :

- des moyens (62) d'analyse dudit signal à convertir (60), regroupé en trames, pour obtenir, pour chaque trame d'échantillons, des informations relatives aux caractéristiques acoustiques ;

- des moyens (64) de détermination d'un indice de correspondance entre les trames à convertir et chaque composante dudit modèle ; et

- des moyens (65) de sélection d'une partie déterminée desdites composantes dudit modèle en fonction desdits indices de correspondance,

lesdits moyens (66) d'application étant adaptés pour appliquer uniquement une partie déterminée de ladite au moins une fonction de transformation correspondant auxdites composantes du modèle sélectionnées.

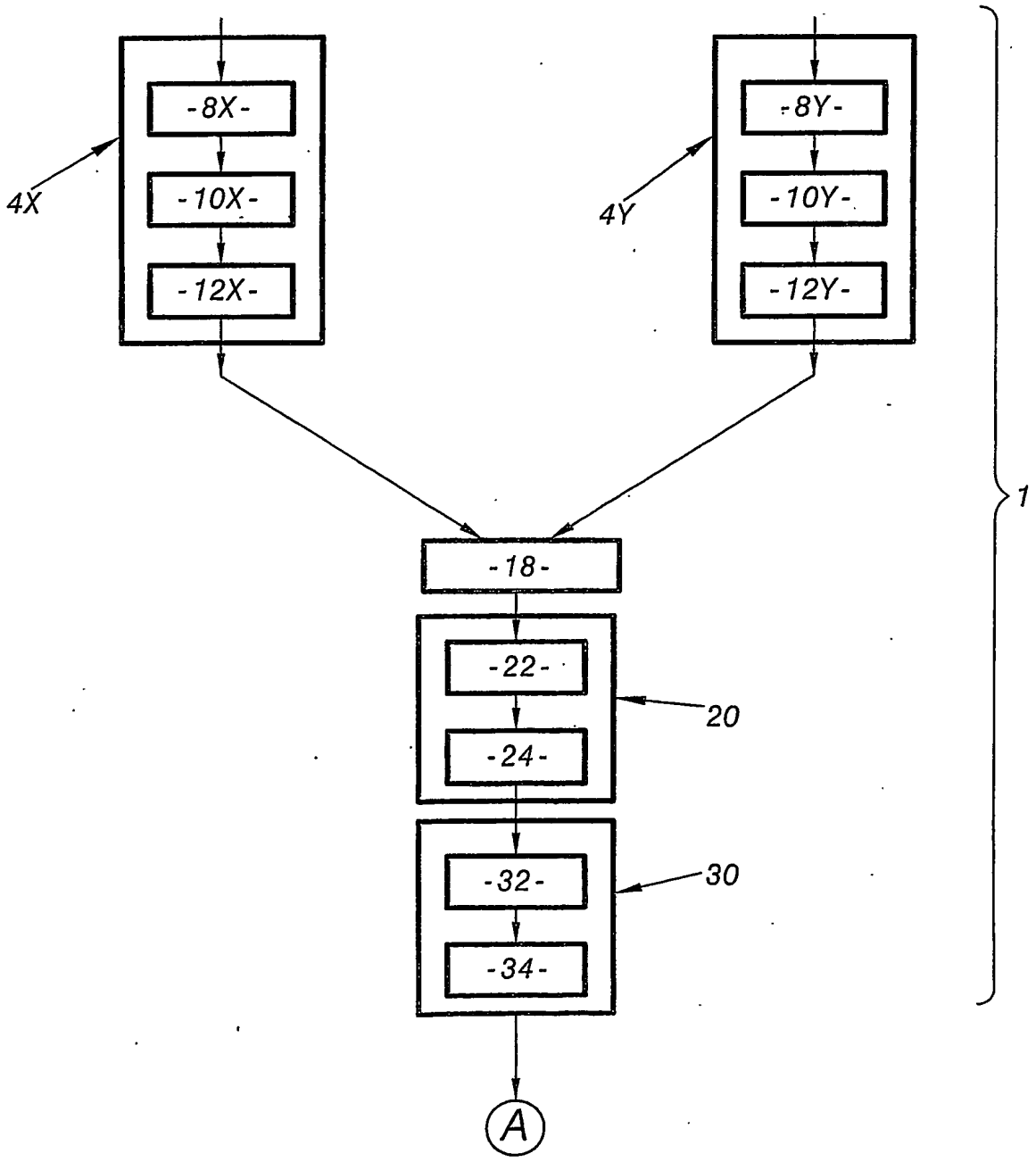

FIG.1A

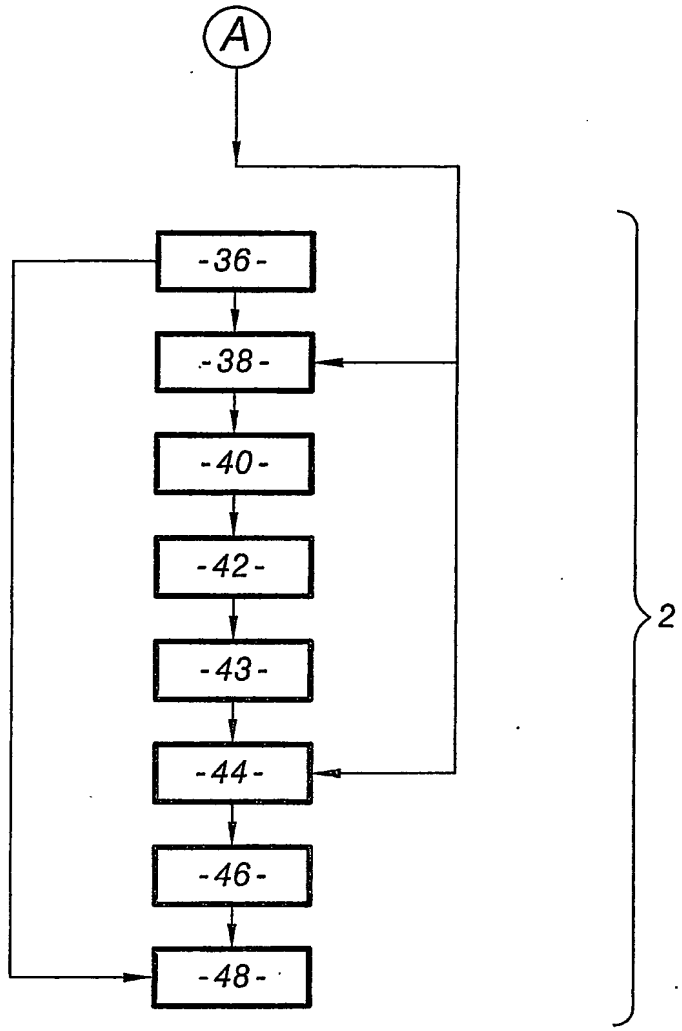

FIG.1B

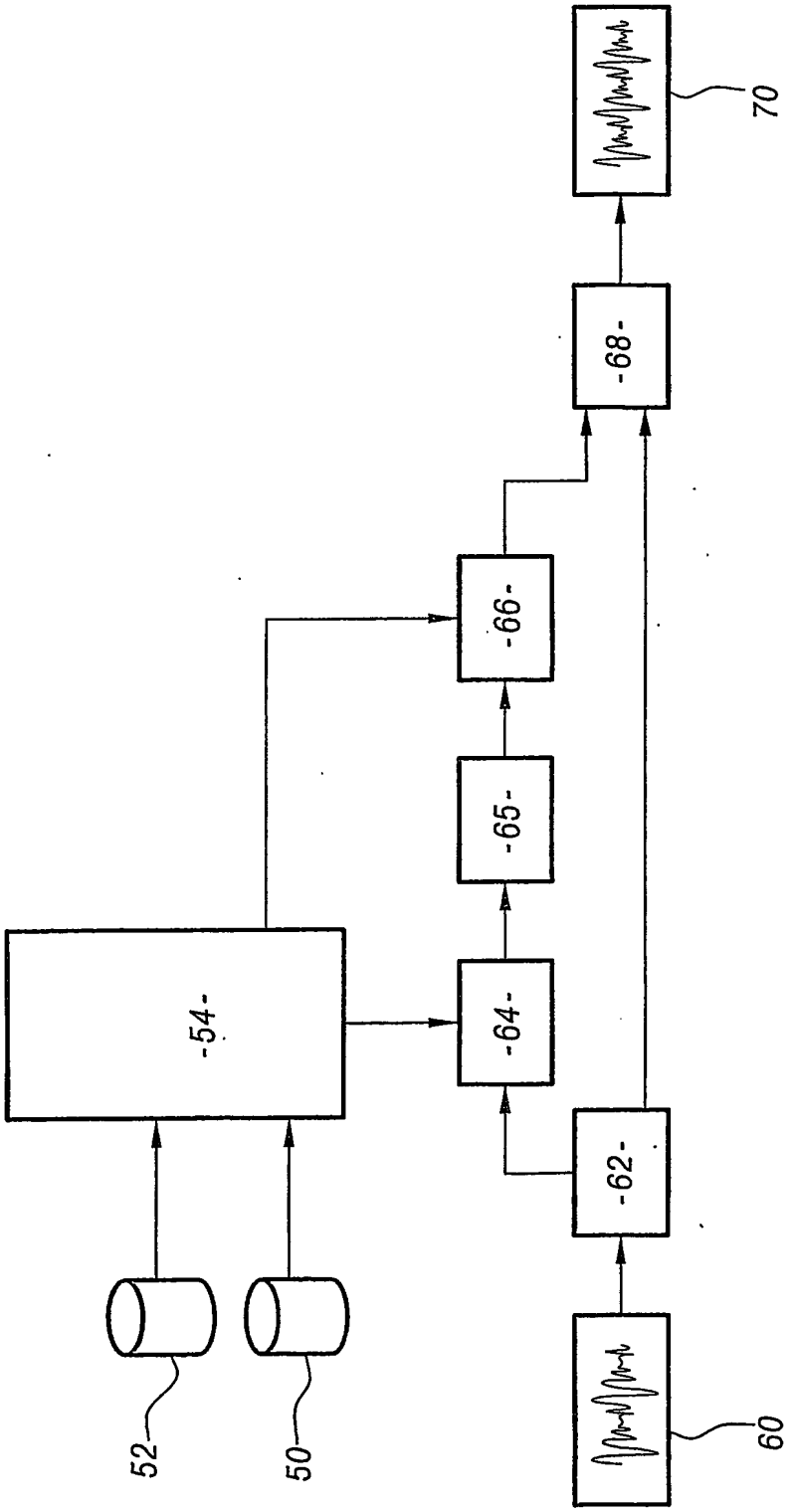

FIG.2

INTERNATIONAL SEARCH REPORT

International Application No
PCT/FR2005/000607

A. CLASSIFICATION OF SUBJECT MATTER
IPC 7 G10L21/00

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC 7 G10L

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)
EPO-Internal, WPI Data, PAJ, IBM-TDB, INSPEC, COMPENDEX

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	STYLIANOU Y ET AL: "STATISTICAL METHODS FOR VOICE QUALITY TRANSFORMATION" 4TH EUROPEAN CONFERENCE ON SPEECH COMMUNICATION AND TECHNOLOGY. EUROSPEECH '95. MADRID, SPAIN, SEPT. 18 - 21, 1995, EUROPEAN CONFERENCE ON SPEECH COMMUNICATION AND TECHNOLOGY. (EUROSPEECH), MADRID : GRAFICAS BRENS, ES, vol. VOL. 1 CONF. 4, 18 September 1995 (1995-09-18), pages 447-450, XP000854745 the whole document ----- -/--	1-12

[X] Further documents are listed in the continuation of box C.

[X] Patent family members are listed in annex.

° Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
24 June 2005	01 09. 2005

Name and mailing address of the ISA	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Tx. 31 651 epo nl, Fax: (+31-70) 340-3016	Dobler, E

Form PCT/ISA/210 (second sheet) (January 2004)

INTERNATIONAL SEARCH REPORT

International Application No
PCT/FR2005/000607

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	WO 02/067245 A (AUCKENTHALER ROLAND ; CAREY MICHAEL JOHN (GB); IMAGINATION TECHNOLOGIE) 29 August 2002 (2002-08-29) page 1, line 34 - page 2, line 16; figures 1-4 page 6, line 10 - page 8, line 9	1-12
A	BANDOIN G ET AL: "On the transformation of the speech spectrum for voice conversion" SPOKEN LANGUAGE, 1996. ICSLP 96. PROCEEDINGS., FOURTH INTERNATIONAL CONFERENCE ON PHILADELPHIA, PA, USA 3-6 OCT. 1996, NEW YORK, NY, USA,IEEE, US, 3 October 1996 (1996-10-03), pages 1405-1408, XP010237945 ISBN: 0-7803-3555-4 page 1405, right-hand column, line 3 - page 1407, left-hand column, line 26	1-12
A	HELENCA DUXANS AND ANTONIO BONAFONTE ET AL: "Estimation of GMM in voice conversion including unaligned data" PROCEEDINGS OF THE EUROSPEECH 2003 CONFERENCE, September 2003 (2003-09), pages 861-864, XP007007125 the whole document	1-12
A	YINING CHEN1 ET AL: "Voice Conversion with Smoothed GMM and MAP Adaptation" PROCEEDINGS OF THE EUROSPEECH 2003 CONFERENCE, September 2003 (2003-09), pages 2413-2416, XP007006960 page 2413, left-hand column, line 1 - page 2415, left-hand column, line 18	1-12
A	LAROCHE J ET AL: "HNM: a simple, efficient harmonic+noise model for speech" APPLICATIONS OF SIGNAL PROCESSING TO AUDIO AND ACOUSTICS, 1993. FINAL PROGRAM AND PAPER SUMMARIES., 1993 IEEE WORKSHOP ON NEW PALTZ, NY, USA 17-20 OCT. 1993, NEW YORK, NY, USA,IEEE, 17 October 1993 (1993-10-17), pages 169-172, XP010130052 ISBN: 0-7803-2078-6 the whole document	1-12

INTERNATIONAL SEARCH REPORT

Information on patent family members

International Application No
PCT/FR2005/000607

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
WO 02067245 A	29-08-2002	GB 2372366 A	21-08-2002
		WO 02067245 A1	29-08-2002

RAPPORT DE RECHERCHE INTERNATIONALE

Demande Internationale No
PCT/FR2005/000607

A. CLASSEMENT DE L'OBJET DE LA DEMANDE
CIB 7 G10L21/00

Selon la classification internationale des brevets (CIB) ou à la fois selon la classification nationale et la CIB

B. DOMAINES SUR LESQUELS LA RECHERCHE A PORTE

Documentation minimale consultée (système de classification suivi des symboles de classement)
CIB 7 G10L

Documentation consultée autre que la documentation minimale dans la mesure où ces documents relèvent des domaines sur lesquels a porté la recherche

Base de données électronique consultée au cours de la recherche internationale (nom de la base de données, et si réalisable, termes de recherche utilisés)
EPO-Internal, WPI Data, PAJ, IBM-TDB, INSPEC, COMPENDEX

C. DOCUMENTS CONSIDERES COMME PERTINENTS

Catégorie °	Identification des documents cités, avec, le cas échéant, l'indication des passages pertinents	no. des revendications visées
Y	STYLIANOU Y ET AL: "STATISTICAL METHODS FOR VOICE QUALITY TRANSFORMATION" 4TH EUROPEAN CONFERENCE ON SPEECH COMMUNICATION AND TECHNOLOGY. EUROSPEECH '95. MADRID, SPAIN, SEPT. 18 - 21, 1995, EUROPEAN CONFERENCE ON SPEECH COMMUNICATION AND TECHNOLOGY. (EUROSPEECH), MADRID : GRAFICAS BRENS, ES, vol. VOL. 1 CONF. 4, 18 septembre 1995 (1995-09-18), pages 447-450, XP000854745 le document en entier ----- -/--	1-12

[X] Voir la suite du cadre C pour la fin de la liste des documents

[X] Les documents de familles de brevets sont indiqués en annexe

° Catégories spéciales de documents cités:

"A" document définissant l'état général de la technique, non considéré comme particulièrement pertinent
"E" document antérieur, mais publié à la date de dépôt international ou après cette date
"L" document pouvant jeter un doute sur une revendication de priorité ou cité pour déterminer la date de publication d'une autre citation ou pour une raison spéciale (telle qu'indiquée)
"O" document se référant à une divulgation orale, à un usage, à une exposition ou tous autres moyens
"P" document publié avant la date de dépôt international, mais postérieurement à la date de priorité revendiquée
"T" document ultérieur publié après la date de dépôt international ou la date de priorité et n'appartenant pas à l'état de la technique pertinent, mais cité pour comprendre le principe ou la théorie constituant la base de l'invention
"X" document particulièrement pertinent; l'invention revendiquée ne peut être considérée comme nouvelle ou comme impliquant une activité inventive par rapport au document considéré isolément
"Y" document particulièrement pertinent; l'invention revendiquée ne peut être considérée comme impliquant une activité inventive lorsque le document est associé à un ou plusieurs autres documents de même nature, cette combinaison étant évidente pour une personne du métier
"&" document qui fait partie de la même famille de brevets

Date à laquelle la recherche internationale a été effectivement achevée	Date d'expédition du présent rapport de recherche internationale
24 juin 2005	01 09. 2005

Nom et adresse postale de l'administration chargée de la recherche internationale	Fonctionnaire autorisé
Office Européen des Brevets, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Tx. 31 651 epo nl, Fax: (+31-70) 340-3016	Dobler, E

RAPPORT DE RECHERCHE INTERNATIONALE

Demande internationale No
PCT/FR2005/000607

C.(suite) DOCUMENTS CONSIDERES COMME PERTINENTS

Catégorie °	Identification des documents cités, avec, le cas échéant, l'indication des passages pertinents	no. des revendications visées
Y	WO 02/067245 A (AUCKENTHALER ROLAND ; CAREY MICHAEL JOHN (GB); IMAGINATION TECHNOLOGIE) 29 août 2002 (2002-08-29) page 1, ligne 34 - page 2, ligne 16; figures 1-4 page 6, ligne 10 - page 8, ligne 9	1-12
A	BANDOIN G ET AL: "On the transformation of the speech spectrum for voice conversion" SPOKEN LANGUAGE, 1996. ICSLP 96. PROCEEDINGS., FOURTH INTERNATIONAL CONFERENCE ON PHILADELPHIA, PA, USA 3-6 OCT. 1996, NEW YORK, NY, USA,IEEE, US, 3 octobre 1996 (1996-10-03), pages 1405-1408, XP010237945 ISBN: 0-7803-3555-4 page 1405, colonne de droite, ligne 3 - page 1407, colonne de gauche, ligne 26	1-12
A	HELENCA DUXANS AND ANTONIO BONAFONTE ET AL: "Estimation of GMM in voice conversion including unaligned data" PROCEEDINGS OF THE EUROSPEECH 2003 CONFERENCE, septembre 2003 (2003-09), pages 861-864, XP007007125 le document en entier	1-12
A	YINING CHEN1 ET AL: "Voice Conversion with Smoothed GMM and MAP Adaptation" PROCEEDINGS OF THE EUROSREECH 2003 CONFERENCE, septembre 2003 (2003-09), pages 2413-2416, XP007006960 page 2413, colonne de gauche, ligne 1 - page 2415, colonne de gauche, ligne 18	1-12
A	LAROCHE J ET AL: "HNM: a simple, efficient harmonic+noise model for speech" APPLICATIONS OF SIGNAL PROCESSING TO AUDIO AND ACOUSTICS, 1993. FINAL PROGRAM AND PAPER SUMMARIES., 1993 IEEE WORKSHOP ON NEW PALTZ, NY, USA 17-20 OCT. 1993, NEW YORK, NY, USA,IEEE, 17 octobre 1993 (1993-10-17), pages 169-172, XP010130052 ISBN: 0-7803-2078-6 le document en entier	1-12

RAPPORT DE RECHERCHE INTERNATIONALE

Renseignements relatifs aux membres de familles de brevets

Demande internationale No
PCT/FR2005/000607

Document brevet cité au rapport de recherche	Date de publication	Membre(s) de la famille de brevet(s)	Date de publication
WO 02067245 A	29-08-2002	GB 2372366 A	21-08-2002
		WO 02067245 A1	29-08-2002